Theoretical Computer Science 81 (1991) 65-75
Elsevier

Probabilistic approach to the satisfiability problem

Olivier Dubois

LAFORIA, CNRS-Université Paris VI (UA 1095), 4 place Jussieu, 75252 Paris Cedex 05, France

Jacques Carlier

Université de Compiègne, UA 81X Heudiasyc, 60200 Compiègne, France

Communicated by M. Nivat
Received May 1988
Revised January 1989

1. Introduction

NP-complete problems are considered intractable because algorithms at our disposal to solve them have an exponential time complexity [3]. This means that the time to compute any solution increases very rapidly above "reasonable" norms as the size of instances increases. Moreover it is conjectured that "good" algorithms with a polynomial time complexity to solve those problems do not exist. Is it possible therefore, to provide a probabilistic answer? We propose a probabilistic approach to the study of NP-complete problems and we report the first step taken towards the use of this approach for the satisfiability problem (SAT problem for short) [1].

In [2] we studied the problem of counting the number of solutions for SAT instances. Here we study this problem from a probabilistic point of view. We first define a probabilistic model of drawing of SAT instances and then we compute the mathematical expectation of the number of solutions. A probabilistic estimate of their contradiction is deduced. We then suggest modalities for a probabilistic study of SAT instances by partitioning the infinite set of instances into finite classes, and by studying the distribution of the number of solutions in each class.

1.1. Notations

Notations used are those mentioned in [2]. We shall assume in the sequel of this paper literals, clauses and SAT instances are formed from the set of variables $X = \{x_1, x_2, \ldots, x_n\}$.

2. Probabilistic model

Let us state the conditions of drawing of SAT instances. A SAT instance is obtained by successive drawings of clauses. A clause is formed by drawing according to a non-defined process a sequence of variables of X, and then drawing the form, either direct or complemented, of each variable of the sequence. The probability law for drawing the form of each variable is a uniform law implying that direct form and complemented form of occurrences of a variable are equally likely.

The sequence of variables of a clause can be considered as an arrangement of variables of X. Let Ω be the set of all possible arrangements of variables of X with possible repetitions. Ω is infinite and represents the set of all possible events resulting from the drawing of a sequence of variables for a clause. For any arrangement $x_{i_1}x_{i_2}\dots x_{i_r}$ with $(i_1, i_2, \dots, i_r)\in[1, n]^r$ and $r\in\mathbb{N}$, a probability denoted by $p_{i_1,i_2,\dots,i_r}$ can therefore be associated. We assume that $r\geq 1$ which excludes drawing of the empty clause (to include it, it requires that $r\geq 0$ and the same reasoning as for $r\geq 1$ follows). By definition, the probability of Ω being equal to 1, we have

$$\sum_{r=1}^{r=\infty} \sum_{i_1,i_2,\dots,i_r\in[1,n]^r} p_{i_1,i_2,\dots,i_r} = 1.$$

Examples. Let us assume that we draw the variables of a clause with a given length r by making r successive independent drawings of a variable in X. Let q_i be the probability of drawing the variable x_i with $i\in[1, n]$. The probability of the arrangement $x_{i_1}x_{i_2}\dots x_{i_r}$ is $q_{i_1}q_{i_2}\cdots q_{i_r}$, and we have

$$\sum_{i_1,i_2,\dots,i_r\in[1,n]^r} q_{i_1}q_{i_2}\cdots q_{i_r} = 1.$$

If one desires clauses with a non-fixed length and if u_r is the probability of drawing a clause with r variables the probability of the previous arrangement becomes $u_r q_{i_1}q_{i_2}\cdots q_{i_r}$. If one desires a clause without repeated variables, the drawing is without replacement and the probability of $x_{i_1}x_{i_2}\dots x_{i_r}$ becomes

$$\frac{q_{i_1}q_{i_2}\cdots q_{i_r}}{(1-q_{i_1})(1-q_{i_1}-q_{i_2})\cdots(1-q_{i_1}-q_{i_2}-\cdots-q_{i_{r-1}})}.$$

One can also obtain variables of a clause by directly drawing arrangements in Ω according to a distribution of probabilities.

Now let us compute the mathematical expectation of the number of solutions of k clauses drawn under the conditions of the probabilistic model which has been defined above.

Firstly let us consider any k clauses having a number of solutions N_k. Let us add a $(k+1)$th clause drawn under the conditions of the model, denoting by ΔN_k the random variable expressing the decrease of N_k when the $(k+1)$th clause is added. Let us assume that the arrangement $x_{i_1}x_{i_2}\dots x_{i_r}$ is drawn for the $(k+1)$th clause. By

complementing the r variables of the arrangement in all possible ways 2^r clauses can be generated (not necessarily all distinct, because variables can be repeated). The direct form and the complemented form being equally likely under the conditions of the probabilistic model, the probability of drawing one of the 2^r clauses generated from the above fixed arrangement is $1/2^r$.

Let s be the number of distinct variables in the arrangement $x_{i_1}x_{i_2}\ldots x_{i_r}$. Among the 2^r clauses generated from this arrangement there are 2^s clauses with s distinct variables and 2^r-2^s clauses which are tautologies. These last clauses do not suppress any solution when added to the considered k clauses. Let $m_1,\ldots,m_t,\ldots,m_{2^s}$ be the number of solutions suppressed by these other 2^s clauses. The conditional expectation of ΔN_k for N_k fixed and the arrangement $x_{i_1}x_{i_2}\ldots x_{i_r}$ fixed is

$$E(\Delta N_k|N_k|x_{i_1}x_{i_2}\ldots x_{i_r})=\sum_{t=1}^{t=2^s}\frac{m_t}{2^r}. \tag{1}$$

Let us recall three definitions stated in [2] and a property.

Definition 2.1. Two clauses non-reducible to tautologies are independent if and only if they have at least one common variable in opposite forms, direct in one and complemented in the other.

Definition 2.2. k clauses are independent if and only if two at a time are independent.

Definition 2.3. If k clauses are independent and if their lengths $r_1,\ldots,r_i,\ldots,r_k$ satisfy the relation, $\sum_{i=1}^{i=k}1/2^{r_i}=1$, these k clauses are base clauses and form a basis.

Property 2.4. Any valuation of variables contradicts one and only one clause or inversely satisfies exactly $k-1$ clauses of a basis.

The 2^s clauses mentioned previously (1) are independent by construction. Moreover $\sum_{i=1}^{i=2^s}1/2^s=1$. Therefore they form a basis.

By Property 2.4 each one of the N_k solutions of the k considered clauses is suppressed by only one of the 2^s clauses. Hence,

$$\sum_{t=1}^{t=2^s}m_t=N_k \quad\text{and}\quad E(\Delta N_k|N_k|x_{i_1}\ldots x_{i_r})=\frac{N_k}{2^r}. \tag{2}$$

Therefore we have

$$E(\Delta N_k|N_k)=\sum_{r=1}^{r=\infty}\sum_{i_1,i_2,\ldots,i_r\in[1,n]^r}\frac{N_k}{2^r}p_{i_1,\ldots,i_r}.$$

Let

$$P_r=\sum_{i_1,i_2,\ldots,i_r\in[1,n]^r}p_{i_1,\ldots,i_r},$$

hence

$$E(\Delta N_k\,|\,N_k)=N_k\sum_{r=1}^{r=\infty}\frac{P_r}{2^r}.$$

Let us now consider N_k, for any $k \in \mathbb{N}$, as a random variable expressing the number of solutions of k clauses drawn under the conditions of the probabilistic model.

We have $N_{k+1} = N_k - \Delta N_k$, hence

$$E(N_{k+1} \mid N_k) = N_k - E(\Delta N_k \mid N_k) = N_k \left(1 - \sum_{r=1}^{r=x} \frac{P_r}{2^r}\right).$$

One deduces

$$E(N_{k+1}) = E(N_k)\left(1 - \sum_{r=1}^{r=x} \frac{P_r}{2^r}\right).$$

Since $E(N_0) = 2^n$, we obtain by induction on k,

$$E(N_k) = 2^n \left(1 - \sum_{r=1}^{r=x} \frac{P_r}{2^r}\right)^k. \tag{3}$$

Equation (3) does therefore express the mathematical expectation of the number of solutions of k clauses drawn under the conditions of the probabilistic model.

Using equation (3) let us investigate the probability that an instance with k clauses satisfying the probabilistic model is unsatisfiable or contradictory. Let us denote by $\text{Prob}(N_k = i)$ the probability that an instance of the model with k clauses has i solutions. $\text{Prob}(N_k = 0)$ is the probability that such an instance is contradictory. We have

$$\sum_{i=0}^{i=2^n} \text{Prob}(N_k = i) = 1,$$

hence,

$$\text{Prob}(N_k = 0) = 1 - \sum_{i=1}^{i=2^n} \text{Prob}(N_k = i).$$

Since

$$\sum_{i=1}^{i=2^n} \text{Prob}(N_k = i) \leq \sum_{i=1}^{i=2^n} i \, \text{Prob}(N_k = i),$$

and

$$\sum_{i=1}^{i=2^n} i \, \text{Prob}(N_k = i) = \sum_{i=0}^{i=2^n} i \, \text{Prob}(N_k = i) = E(N_k),$$

then $\text{Prob}(N_k = 0) \geq 1 - E(N_k)$. Let

$$a = 2^n \left(1 - \sum_{r=1}^{r=x} \frac{P_r}{2^r}\right)^k. \tag{4}$$

We can state that for any instance with k clauses drawn under the conditions of the probabilistic model, the probability that the instance is contradictory is greater than or equal to $1 - a$ or $\text{Prob}(N_k = 0) \geq 1 - a$ (obviously this result is interesting only when a approaches 0).

This probabilistic evaluation gives a lower bound of the probability of contradiction. We carried out experiments to compare theoretical values provided by (4) with experimental values.

Experiments consisted of successively drawing random clauses with three distinct variables until a contradictory instance was obtained. For a given number of contradictory instances drawn we determined the number k of clauses of these instances for which the contradiction occurred in half the instances. We compared the values of the ratio k/n with those provided by (4) for $a = \frac{1}{2}$. Under the conditions of experiments $P_r = 1$ for $r = 3$ and $P_r = 0$ for any other value of r. Hence,

$$a = 2^n \left(1 - \frac{1}{2^3}\right)^k,$$

and

$$k/n = -\frac{\ln(2)}{\ln(7/8)}\left(1 + \frac{1}{n}\right) \quad \text{for } a = \tfrac{1}{2}. \tag{5}$$

Table 1 gives the results of experiments.

The probabilistic calculation provides information about the set of instances which can be drawn. In the next section classes of instances will be defined as small as possible such that a probabilistic calculation can be applied. To cover the set of all SAT instances, classes must partition this set.

Table 1

n number of variables	k/n	
	Values provided by (5)	Experimental values
10	5.71	5.10
20	5.45	5.35
30	5.36	4.77
40	5.32	4.68
50	5.29	4.62
60	5.28	4.51
70	5.27	4.44
80	5.26	4.41
90	5.25	4.40
100	5.24	4.43

3. Partition of SAT into classes of instances and probabilistic calculation

Let us term "structure" of an instance the distribution of lengths of clauses of this instance, that is the set S of couples (k_i, r_i), k_i is the number of clauses with

length r_i for $i \in [1, p]$ and $K = \sum_{i=1 \text{ to } p} k_i$ the total number of clauses of the instance. One denotes $S = \{(k_i, r_i)_{i \in [1,p]}\}$.

Let us term "distribution of variables" of an instance in the structure S of this instance, the set D of combinations of variables such that each combination is the set of variables of a clause of the instance.

Example. For the following SAT instance:

$$\bar{x}_4 \vee x_6, \qquad x_1 \vee \bar{x}_2 \vee x_3, \qquad \bar{x}_2 \vee x_4 \vee x_5,$$

we have $S = \{(1, 2), (2, 3)\}$ and $D = \{(x_4\, x_6), (x_1\, x_2\, x_3), (x_2\, x_4\, x_5)\}$.

From a combination with r variables, 2^r clauses with a length r (not necessarily all distinct) can be generated by complementing variables in all possible ways. From a structure $S = \{(k_i, r_i)_{i \in [1,p]}\}$ and a distribution of variables D in S, $2^{\sum_{i=1 \text{ to } p} k_i r_i}$ instances (not necessarily all distinct) can be generated by complementing variables of D in all possible ways. Let us regroup these instances with the structure S and the distribution of variables D in a class denoted by $C(S, D)$. For all possible sets S and D, classes $C(S, D)$ partition the infinite set of SAT instances.

Theorem 3.1. *Let X be a set of n variables and $C(S, D)$ a class of instances with $S = \{(k_i, r_i)_{i \in [1,p]}\}$ and D a distribution of variables of X in S. The mean of the numbers of solutions of the $2^{\sum_{i=1 \text{ to } p} k_i r_i}$ instances of $C(S, D)$ is*

$$N_{C(S,D)} = 2^n \prod_{i=1}^{i=p} \left(1 - \frac{1}{2^{r_i}}\right)^{k_i}.$$

Proof. Let $C_1, \ldots, C_K$ be the combinations of D having $r_1, \ldots, r_K$ variables, respectively, with $K = \sum_{i=1 \text{ to } p} k_i$. Let us consider j clauses formed from the first j combinations of D: $C_1, \ldots, C_j$, and let N_j be their number of solutions. Let us use probabilistic reasoning of Section 2. Let us add a $(j+1)$th clause to the considered first j clauses using a probability law of drawing of variables of the $(j+1)$th clause such that the combination C_{j+1} of D has the probability 1 to be drawn, and using a uniform law of drawing of the form of each variable of C_{j+1}.

We have by (2), $E(\Delta N_j | N_j | C_{j+1}) = N_j / 2^{r_{j+1}}$. The probability of drawing C_{j+1} being 1, $E(\Delta N_j | N_j) = N_j / 2^{r_{j+1}}$. Hence,

$$E(N_{j+1} | N_j) = N_j \left(1 - \frac{1}{2^{r_{j+1}}}\right).$$

Let us draw an instance with K clauses by successively drawing each one of the combinations of D with a probability of 1, and drawing the forms of each variable according to a uniform law. We have for $j+1$ clauses drawn

$$E(N_{j+1}) = E(N_j) \left(1 - \frac{1}{2^{r_{j+1}}}\right). \tag{6}$$

By induction on j we obtain from (6),

$$E(N_k) = 2^n \prod_{j=1}^{j=K} \left(1 - \frac{1}{2^{r_j}}\right),$$

or by regrouping the lengths r_j according to the structure S,

$$E(N_K) = 2^n \prod_{i=1}^{i=p} \left(1 - \frac{1}{2^{r_i}}\right)^{k_i}.$$

The set of instances which can be drawn under the conditions fixed above is the class $C(S, D)$. Moreover these instances have an equal probability of being drawn. Therefore $E(N_K)$ is the mean of the numbers of solutions of instances in the class $C(S, D)$. □

As in Section 2 one derives from the mean value $N_{C(S,D)}$ a probabilistic estimate of the contradiction of a random instance in $C(S, D)$.

Corollary 3.2. *If $C(S, D)$ is a class of instances, the probability that a random instance drawn from $C(S, D)$ is contradictory is greater than or equal to $1 - N_{C(S,D)}$ (obviously this result is interesting only when $N_{C(S,D)}$ approaches 0).*

The main consequence of Theorem 3.1 is that the mean number of solutions of a class $C(S, D)$ is independent of the distribution D of variables, but is dependent only on the structure S. For a given structure S_0 all classes $C(S_0, D)$ have the same mean number of solutions. Let us take as an extreme example the distribution δ in S_0 which contains only one of the n variables of X, this one being repeated as many times as necessary. The instances of the class $C(S_0, \delta)$ have 2^n or 2^{n-1} or 0 solutions over X. The mean of these numbers of solutions is the same as for any other distribution D and is equal to $N_{C(S_0,D)}$.

However classes $C(S_0, D)$ differ in the dispersion of the numbers of solutions, the dispersion depending on the distribution of variables D. We intend to study this dispersion. But the following preliminary remarks can be made.

Dispersion of the numbers of solutions in $C(S, D)$ can be studied in the quotient space $C(S, D)/R$ where R is the following equivalence relation: an instance A is in the relation R to an instance A' if and only if $A' = A$ or A' can be derived from A by complementing all occurrences of one or more variables of A. Any two instances in the relation R have the same number of solutions. If n is the number of distinct variables in D any equivalence class contains 2^n instances of $C(S, D)$. Let $S = \{(k_i, r_i)_{i\in[1,p]}\}$, then there are $2^{(\sum_{i=1 \text{ to } p} k_i r_i) - n}$ equivalence classes in $C(S, D)$. $C(S, D)$ and $C(S, D)/R$ have the same distribution of the probabilities of the numbers of solutions.

We are trying to characterize instances in classes $C(S, D)$ which correspond to the limits of the dispersion. Below we characterize those having the maximum of solutions. Firstly let us establish the following proposition.

Proposition 3.3. *Let A be a SAT instance over a set X of variables without repeated variables in a clause, and let x_i be a variable of X appearing in A. The instance A' derived from A by changing the forms of all occurrences of x_i into only one form has a number of solutions over X greater than or equal to that of A. We write $N(A') \geq N(A)$.*

Proof. Let K be the number of clauses of A. Let us prove the proposition by induction on K. For $K = 1$ the proposition is true. Let us assume that the proposition holds for any instance with $K-1$ clauses. Let A be an instance with K clauses denoted by $C_1, C_2, \ldots, C_j, \ldots, C_K$. Let us denote by $C'_1, C'_2, \ldots, C'_j, \ldots, C'_K$ clauses of A' derived from A. When x_i appears in A in only one form, direct or complemented, the proposition is true. Now let us consider the case when x_i appears in A in both forms. Without loss of generality we assume that the form of occurrences of x_i is changing into the direct form in A', and that clause C_K contains the literal x_i. We denote

$$N(A) = N\left[\bigwedge_{j=1}^{j=K} C_j\right] \quad \text{and} \quad N(A') = N\left[\bigwedge_{j=1}^{j=K} C'_j\right].$$

We have established in [2] the following lemma.

Lemma 3.4. *Let F_1 and F_2 be two boolean formulas over a set X of variables, we then have the relation $N(F_1 \wedge F_2) = N(F_1) + N(F_2) - N(F_1 \vee F_2)$.*

From this lemma we can write

$$N\left[\bigwedge_{j=1}^{j=K} C_j\right] = N\left[\left(\bigwedge_{j=1}^{j=K-1} C_j\right) \wedge C_K\right]$$

$$= N\left[\bigwedge_{j=1}^{j=K-1} C_j\right] + N\left[C_K\right] - N\left[\left(\bigwedge_{j=1}^{j=K-1} C_j\right) \vee C_K\right].$$

By induction hypothesis

$$N\left[\bigwedge_{j=1}^{j=K-1} C_j\right] \leq N\left[\bigwedge_{j=1}^{j=K-1} C'_j\right].$$

We have $N[C_K] = N[C'_K]$.

Let us consider the two following sets:

$$E = \{(C_1 \vee C_K), (C_2 \vee C_K), \ldots, (C_j \vee C_K), \ldots, (C_{K-1} \vee C_K)\},$$

$$E' = \{(C'_1 \vee C'_K), (C'_2 \vee C'_K), \ldots, (C'_j \vee C'_K), \ldots, (C'_{K-1} \vee C'_K)\}.$$

We have $C_K = C'_K$. For each clause C_j with $j \in [1, K-1]$, one of the following three cases occurs:

- C_j does not contain the variable x_i, then $C_j \vee C_K = C'_j \vee C'_K$;
- C_j contains the literal x_i, then $C_j \vee C_K = C'_j \vee C'_K$;
- C_j contains the literal $\bar{x}_i$, then $C_j \vee C_K$ is equivalent to a tautology.

Consequently,

$$N\left[\left(\bigwedge_{j=1}^{j=K-1} C_j\right) \vee C_K\right] \geqslant N\left[\left(\bigwedge_{j=1}^{j=K-1} C'_j\right) \vee C'_K\right].$$

Thus

$$N\left[\bigwedge_{j=1}^{j=K} C_j\right] \leqslant N\left[\bigwedge_{j=1}^{j=K-1} C'_j\right] + N[C'_K] - N\left[\left(\bigwedge_{j=1}^{j=K-1} C'_j\right) \vee C'_K\right],$$

or

$$N\left[\bigwedge_{j=1}^{j=K} C_j\right] \leqslant N\left[\bigwedge_{j=1}^{j=K} C'_j\right],$$

and $N(A) \leqslant N(A')$. □

Now let us characterize instances of classes $C(S, D)$ having the maximum number of solutions. Let us assume that each combination of D is without repetition of variables. Otherwise, a clause is formed from each combination of D containing repeated variables such that for two occurrences of a variable one is put in the direct form and the other in the complemented form. These clauses are equivalent to a tautology which by definition does not suppress any solution. We are brought back to the previous case.

Proposition 3.5. *Let $C(S, D)$ be a class of instances without repetition of variables within each combination of D. Instances of $C(S, D)$ having all occurrences of each variable in only one form (direct or complemented) have the maximum of solutions in $C(S, D)$.*

Proof. Let M be an instance of $C(S, D)$ satisfying the condition of Proposition 3.5. Let n be the number of distinct variables in D. With any instance Z of $C(S, D)$ one can associate a sequence of instances $(Z_0, Z_1, \ldots, Z_n)$ with $Z_0 = Z$ and Z_i obtained from Z_{i-1} by changing the forms of all occurrences of the ith variable of Z into its form in M. Then $Z_n = M$ and by Proposition 3.3 $N(Z_0) \leqslant N(Z_1) \leqslant \cdots \leqslant N(Z_n)$. □

Finally let us identify classes $C(S, D)$ with zero dispersion.

Proposition 3.6. *Classes $C(S, D)$ with zero dispersion are those for which each variable of D has only one occurrence.*

Proof. Let $S = \{(k_i, r_i)_{i \in [1,p]}\}$, and n the number of distinct variables in D. If the number of clauses $K = \sum_{i=1 \text{ to } p} k_i$ equals 1 the proposition is trivial. We assume that $K > 1$.

Sufficiency. If each variable in D has only one occurrence, then $\sum_{i=1 \text{ to } p} k_i r_i = n$. The number of equivalence classes modulo R in $C(S, D)$ is $2^{(\sum_{i=1 \text{ to } p} k_i r_i) - n} = 1$. Therefore all instances have the same number of solutions equal to $N_{C(S,D)}$.

Necessity. We shall show that if at least one variable in D has more than one occurrence, at least two instances of $C(S, D)$ exist having different numbers of solutions. Consequently the dispersion cannot be zero.

Let M be the instance of $C(S, D)$ such that all occurrences of all variables are in the direct form. Let us consider two cases.

Case 1. No clause of M is included in or is equal to another. Inclusion between two clauses means that the set of literals of one is included in the set of literals of the other. Let us consider two subcases.

Subcase 1. There exists some variable x which is repeated in a clause C of M. The instance M' derived from M is formed by complementing one occurrence of x in C. The clause C so transformed is a tautology in M'. Consequently, any solution of M is a solution of M'. Let us consider the valuation such that all variables of X have the value 1 except those belonging to C which have the value 0. This valuation is a solution of M' since by hypothesis no clause of M' is included in or equal to C and therefore any clause of M' has at least one variable distinct from variables of C. This valuation is not a solution of M. Hence $N(M') > N(M)$. M and M' have different numbers of solutions.

Subcase 2. The variables within each clause of M are distinct. We then consider two clauses C_1 and C_2 of M containing a same variable x. The instance M' derived from M is formed by replacing C_2 by C_2' which contains all variables of C_2 in complemented form.

All solutions of M are solutions of M' except those for which all variables of C_2 have the value 1. Inversely, all solutions of M' are solutions of M except those for which all variables of C_2' have the value 0. Let m be the reduced instance of M by assigning the value 1 to all variables of C_2 and let m' be the reduced instance of M' by assigning the value 0 to all variables of C_2'. m is satisfiable by definition of M. m' is satisfiable by hypothesis since no clause of M is included in or equal to C_2. Let us denote by $N(m)$ and $N(m')$ the number of solutions of m and m' respectively over the set of variables of D excluding variables of C_2. From what precedes, it follows therefore that $N(M) - N(m) = N(M') - N(m')$. m' contains all clauses of m plus at least the reduced clause C_1 denoted C_1'. Thus any solution of m' is a solution of m. On the contrary, the valuation such that all variables of m' have the value 1, except those belonging to C_1' which have the value 0, is not a solution of m' but is a solution of m, no clause of M being included in C_1 and therefore C_1'. Then $N(m') < N(m)$ and $N(M') < N(M)$.

Case 2. There exists inclusion or equality relations between some clauses of M. The instance M' derived from M is formed by removing from two clauses in inclusion relation the one with the greatest length, or any one of two clauses in equality relation. We have $N(M') = N(M)$. Let E be the set of clauses removed from M. M' satisfies the condition of the first case. Thereby an instance M'' belonging

to the same class of instances as M' can be formed such that $N(M'') \neq N(M')$ ($>$ or $<$ according to the subcase). Clauses of M'' are those of M' with possible changes in the forms of variables. Let us add all clauses of E to M'' changing the form of variables in such a way that at least one clause of M'' is included in or equal to each clause from E. We obtain an instance M''' belonging to the same class of instances as M, $N(M''') = N(M'')$ and $N(M) \neq N(M''')$. □

4. Conclusion

The basic idea is to exchange certitude for complexity. In this paper we have started a probabilistic study of the SAT problem.

The set of SAT instances has been partitioned into classes defined by a structure of instances and a distribution of variables in the structure. Distribution of instances in these classes is studied according to the number of their solutions. The mean of the numbers of solutions in classes has been determined and we have shown that it is independent of the distribution of variables. A lower bound of the probability of contradiction for a random instance drawn in a class has been given. The study of dispersion of the numbers of solutions in classes is pursued. Moreover this study can improve our understanding of the difficulty in solving SAT instances.

References

[1] S.A. Cook, The complexity of theorem-proving procedures, in: *Proc. 3rd Ann. ACM Symp. on Theory of Computing* (Assoc. Comput. Mach., New York, 1971) 151-158.

[2] O. Dubois, Counting the number of solutions for instances of satisfiability, *Theoret. Comput. Sci.* **81** (1991) 49-64.

[3] M.R. Garey and D.S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-completeness* (Freeman, San Fransisco, 1979).

[4] J.C. Simon, J. Carlier, O. Dubois and O. Moulines, Etude statistique de l'existence de solutions de problèmes SAT, *C.R.A.S. Paris* **302** (I, 7) (1986) 283-286.

[5] J.C. Simon and O. Dubois, Number of solutions of SAT-instances, in: *E.S.P.R. Workshop* Novosibirsk (1987); *I.J.P.R.A.I.* 3(1) (1989) 53-65.